Supplementary Datasets for “Co-citations In Context"

James Bradley, Sitaram Devarakonda, Avon Davey, Siyu Liu, Dmitriy Korobskiy, Tandy Warnow, George Chacko

Corresponding author: George Chacko, E-mail: netelabs@nete.com

Supporting Information Text

This document provides supplemental data sets for an article that is under review. Where data sets are larger than a single page, a link to a Github is provided. All Python, R, SQL, Spark, and Bash code are archived on our Github site. https://github.com/NETESOLUTIONS/ERNIE/tree/master/P2_studies

Relevant folders at this level are

- cocitation_analysis- code for statistical analysis of MCMC data
- imm_metabolism- data sets generated by high level search terms
- workflow_pnas.001.jpeg- graphic of our principal MCMC workflow (jpeg)
- workflow_pnas.pdf- graphic of our principal MCMC workflow (pdf)
- workflow_png.001.png- graphic of our principal MCMC workflow (png)
- Permutation_Testing- assorted production and analytical scripts in SQL, R, Python, Bash

Discussion of algorithmic approach in Uzzi et al (2013) vs the one we used. The MCMC algorithmic approach in Uzzi et al (2013), DOI: 10.1126/science.1240474 for citation switching involves building three dicts containing publications, references, and year of publication information, and using them as lookup tables for various operations. In plain language, an iteration process selects publication in turn. Then each reference in said publication is replaced by a random selection from the *set* of eligible references published in the same year. If the potential replacement candidate is not the same as the reference to be replaced then a replacement is made. If it is the same, then up to 20 tries are made to find a non-self replacement. This process occurs for all the references in the set of publications being analyzed. Thus, reference a in three publications A,B,C could be replaced by references [b,c,d] or [b,b,d] but not [a,b,c]. Secondly, a reciprocal switch is made with a publication that cites the replacement. Thus, if publication A cites reference a published in year X then a is substituted with reference b also published in year X and and a randomly selected publication, say publication B that cites b, will have b replaced with a. See 'satyam_mukherjee_mcmc.py' kindly provided by the authors of this paper DOI: 10.1126/science.1240474.

Our approach is roughly similar. References are first grouped by year of publication and then the sample function in R is used on the *multi-set* of potential replacements to permute all references in a single step. A check is then run to see if the permutation process has created any duplicate references within each publication. Those publications with duplicate references are then deleted (typically <= 0.2%). See 'permute_script.R'.

A key difference is that the pool of replacement candidates is the *set* in one case and the *multi-set* in the other. Every substitution in the first approach is independent for instances of the same reference. Using the *multi-set* accounts for existing citation frequency when selecting possible replacements. Thus a publication in year X that has accumulated 10,000 citations is more likely to be selected than a publication that is cited only once. Reference a in publications A,B,C could be replaced by references [a,b,c]. This process is very fast in comparison and we have scaled it up even further by porting it to the Spark environment. In a recent comparison of publications in WoS in year 1985 (39,1860 pubs and 5,588,861 total references), ten simulations using the satyam_mukherjee_mcmc.py code took roughly 22 hours per simulation on a 32 Gb CentOS VM. Run times using our approach on comparable hardware amounted to roughly 30 seconds per shuffle plus an overhead of 3-4 hours to consolidate the data. We also performed 1,000 simulations in 60 hrs using a small Spark cluster for the much large 2005 data set (886,648 publications and 19,036,324 total references)

For input data we selected all publications of type 'Article' in WoS for a given year. Articles are then filtered to those that have at least two references in them. Further, only those references that have complete records in the Web of Science Core Collection are considered. This eliminates those that have cryptic references to other data sources or are just placeholders. Publications and references are mapped to their respective journals using ISSNs as identifiers. Where a reference has more than one ISSN, the most popular one is assigned to ensure that each reference is associated only with one journal.

For n <= 1000 simulations on disciplinary networks (immunology, metabolism, applied physics) permute_script.R is used to generate n files each with shuffled references. Typically, run times are less than 2 min per simulation on a 32 Gb CentOS box with 8 vCPUs in MS Azure. The permutation_testing_script.sh shell script is then run, which calls four Python scripts in turn.

1. observed_frequency.py: generates journal pair frequencies for the year slice of the WoS or disciplinary data set being analyzed.
2. background_frequency.py: generates journal pair frequencies for the background model implemented using permute_script.R
3. journal_count.py: joins all permuted files generated by background_frequency and calculates mean,std and z-scores.
4. Table_generator.py: final output file which contains all publications, reference pairs along with observed frequency and z-scores.

Thus, the workflow is (i) generate year slice (input data) (ii) generate background models by shuffling references (iii) calculate journal pair frequencies (iv) consolidate observed and simulated frequencies into a single table and calculate z-scores.

This process tends to slow down with large data set such as WoS in 2005 with 886,000 publications and 5.8 million journal pairs. Consequently, the entire process has been ported to Spark and provisioning a cluster, copying source data from the ERNIE PostgreSQL database over to Spark, conducting in-memory calculations, and copying a final table back to PostgreSQL has been automated (see Spark folder). Comparative performance data has been generated and will be posted soon.

Table S1. Counts of Publications and References used in this study

	Year	unique publications	unique references	total references
1	1985	391860	2266584	5588861
2	1986	402309	2316451	5708796
3	1987	412936	2427347	5998513
4	1988	426001	2545647	6354917
5	1989	443144	2673092	6749319
6	1990	458768	2827517	7209413
7	1991	477712	2977784	7729776
8	1992	492181	3134109	8188940
9	1993	504488	3278102	8676583
10	1994	523660	3458072	9255748
11	1995	537160	3680616	9875421
12	1996	663110	4144581	11641286
13	1997	677077	4340733	12135104
14	1998	693531	4573584	12728629
15	1999	709827	4784024	13280828
16	2000	721926	5008842	13810746
17	2001	727816	5203078	14261189
18	2002	747287	5464045	15001390
19	2003	786284	5773756	16024652
20	2004	826834	6095594	17167347
21	2005	886648	6615824	19036324

These data were generated using the uz_ds.sql script in (1). Only publications of type Article and references for which complete records exist in the Web of Science Core Collection are included. The data set consists of 138.6 million unique references that are cited 398.9 million times by 19.7 million publications spanning 21 years (1985-2005)

Table S2. Kullback-Leibler Distances between simulated and observed frequencies.

	data set	D1	D2	D3
1	ap_2005	1.13	0.95	1.04
2	wos_2005 (ap)	1.80	2.35	2.08
3	imm_2005	0.82	0.73	0.78
4	wos_2005 (imm)	1.71	1.92	1.81
5	metab_2005	1.25	1.19	1.22
6	wos_2005 (metab)	2.12	2.60	2.36
7	ap_1995	0.89	0.86	0.87
8	wos_1995 (ap)	1.82	2.37	2.10
9	imm_1995	0.85	0.78	0.81
10	wos_1995 (imm)	1.56	1.70	1.63
11	metab_1995	1.10	1.07	1.22
12	wos_1995 (metab)	1.91	2.33	2.12
13	ap_1985	1.22	1.21	1.22
14	wos_1985 (ap)	1.96	2.37	2.17
15	imm_1985	0.82	0.75	0.79
16	wos_1985 (imm)	1.56	1.68	1.62
17	metab_1985	1.15	1.11	1.13
18	wos_1985 (metab)	1.91	2.24	2.08

The Kullback-Leibler (K-L) Divergence was calculated for simulated (mean value from 1000 simulations) and observed journal pair frequencies for the set of journal pairs common to a disciplinary network and the WoS network, e.g., ap_2005 and wos_2005 (ap). Journal pair frequencies were converted to a probability distribution and rhe *seewave* package in R and K-L_distance_1985.R, K-L_distance_1995.R, and K-L_distance_2005.R scripts were used. D1 and D2 are asymmetric distances, D3 is the symmetric K-L distance.

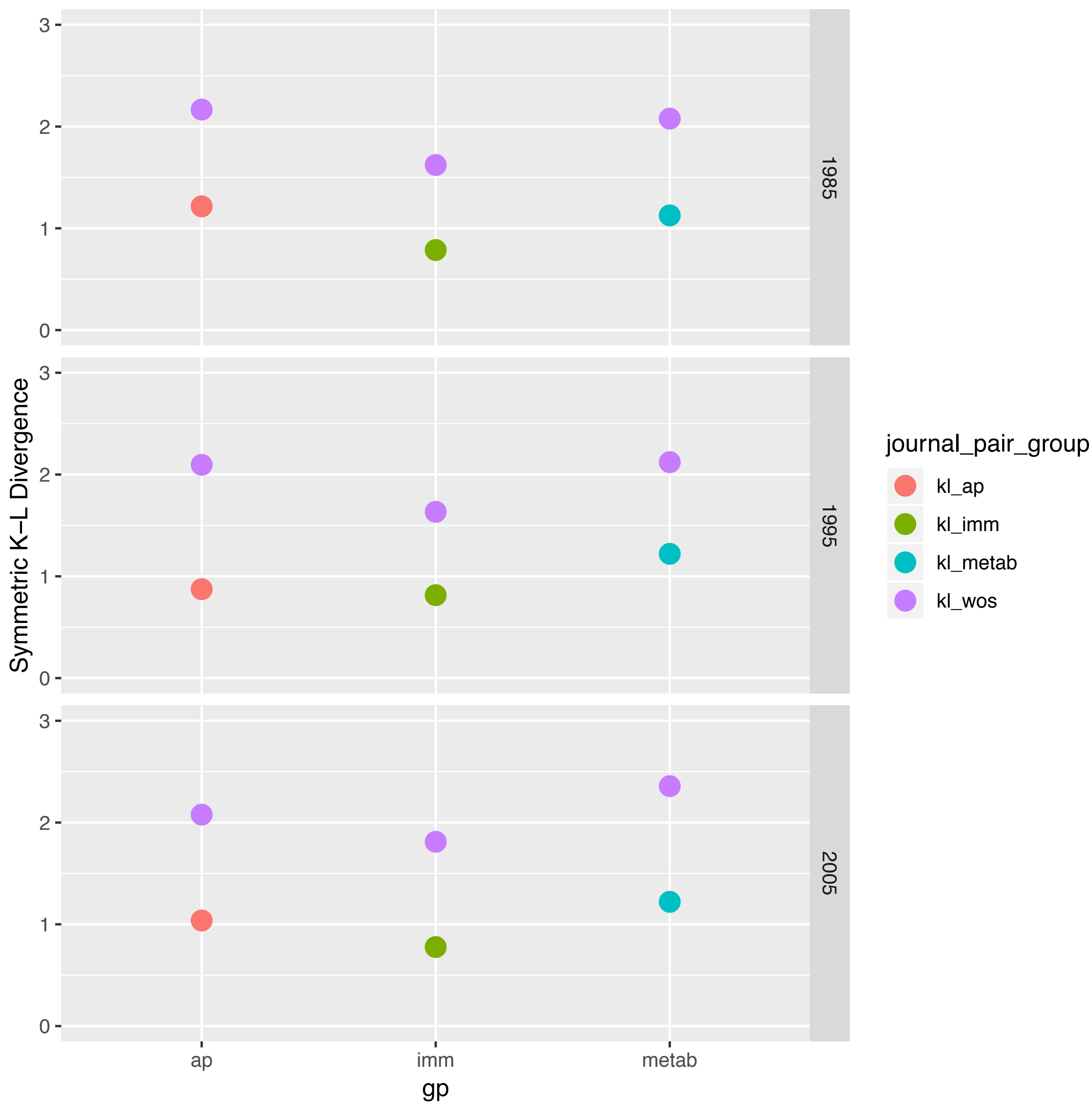

Fig. S1. K-L Divergence demonstrates improved model fits to observed journal pair frequencies for disciplinary networks compared to WoS. For the same journal pairs. the K-L divergence for the WoS network is consistently greater than that for the disciplinary network. from Table S3 are plotted for the years 1985, 1995, and 2005. kl_app, kl_imm, kl_metab, and kl_WoS refer to the symmetrical Kullback-Leibler divergence for mean simulated frequencies and observed frequencies in each of these data sets. The plot is faceted by year.

Table S3. Comparison of Citation Switching Algorithms

	data set	Wos 1985	WoS 1985	WoS 1985	WoS 1995	WoS 2005
1	Input publications	391,860	391,860	391,860	537,160	886,648
2	Input journals	8,075	8,075	8,075	10,983	15,203
3	Observed input journal pairs	1,277,349	1,277,349	1,277,349	2,373,226	5,847,432
4	Simulated journal pairs	961,487	959765	1,200,403	2,288,225	5,835,794
5	Journal pair coverage	75.27%	75.14%	93.97%	96.41%	99.80%
6	Min z-score	-132.71	-148.14	-104.50	-215.96	-273.708
7	Q1 z-score	-2.131	-2.151	-1.43	-1.49	-1.56
8	Median z-score	-0.536	-0.54	-0.24	-0.25	0.555
9	Q3 z-score	3.333	3.365	4.29	4.15	2.423
10	Max z-score	16598.534	22015.891	12,028.55	12,662.15	6,152.57
11	Environment	CentOS 7.4	Spark 2.3	Spark 2.3	Spark 2.3	Spark 2.3
12	Number of simulations	10	10	1000	1000	1000
13	Run time	2186h (22 hr /sim)	$<$ 1 hr	$<$ 50h	50h	60h
14	Algorithm	(2)	(1)	(1)	(1)	(1)

The citation switching algorithm of Uzzi et al. (2013) (2) has been implemented in Python. Ten simulations of the WoS 1985 data set were executed on a 32 Gb, 8 vCPU CentOS 7.4 virtual machine to generate the data in Col 2. Each simulation took roughly 22 hours to complete. Scaling up the experiment, 10 or 1000 simulations of our modifications (1) of this algorithm were executed on a 4-node Apache Spark cluster for the Wos 1985 data set. 1000 simulations completed in less than 50 hours (roughly 3 minute/simulation). These simulations also produced greater journal pair coverage. Performance was compared to 10 or 1000 simulations of our modifications of this algorithm (1) on a 9 node Spark cluster. 1000 simulations completed in less than 50 hours with greater journal pair coverage.

Table S4. Profile of Disciplinary Networks

	data set	Input Publications	Journal Pairs	Min	Q1	Median	Q3	Max
1	ap1985	10298	34,267	-23.05	-0.94	-0.21	3.03	1490.42
2	ap1995	21012	60,340	-45.36	-0.97	-0.24	2.86	646.03
3	ap2005	35600	199,928	-47.76	-0.80	-0.20	3.53	2158.47
4	imm85	17942	159,107	-48.33	-1.09	-0.27	2.49	934.63
5	imm95	22759	319,855	-59.56	-1.10	-0.28	2.37	1507.61
6	imm2005	28539	751,950	-74.54	-0.99	-0.30	1.84	2560.51
7	metab1985	67342	431,993	-97.00	-1.46	-0.34	2.34	4193.49
8	metab1995	100350	865,406	-132.85	-1.56	-0.37	2.16	3998.44
9	metab2005	159910	2,349,005	-127.81	-1.60	-0.41	1.83	3472.77

Data shown represent the results of 1000 simulations for the applied physics (ap), immunology (imm), and metabolism (metab disiplinary network. Summary statistics for z-scores are provided as well as the number of publications in each data set that were the input to the simulation process.

Table S5. Comparison of top 5% cited publications vs all publications in applied physics (ap) immunology (imm), metabolism (metab), and WoS data sets

	year	category	ap_5	ap_all	imm_5	imm_all	metab_5	metab_all	wos_5	wos_all
1	1985	HCHN	34	25	29	33	24	29	6	6
2	1985	HCLN	15	25	21	17	26	21	44	44
3	1985	LCHN	50	47	48	49	48	48	32	29
4	1985	LCLN	1	3	2	1	2	2	18	21
5	1995	HCHN	35	27	29	34	26	31	6	7
6	1995	HCLN	15	23	21	16	24	19	44	43
7	1995	LCHN	50	48	48	49	48	48	33	29
8	1995	LCLN	0	2	2	1	2	2	17	21
9	2005	HCHN	31	29	36	36	30	30	8	7
10	2005	HCLN	19	20	14	14	20	20	42	43
11	2005	LCHN	48	49	50	49	49	49	30	27
12	2005	LCLN	2	2	0	1	1	1	20	23

Numbers shown are percent of publications in each group. Data are shown for reference years 1985, 1995, and 2005

Table S6. Statistical Significance of Deviation from a Random Distribution of Hits

Data Set	Year	Highly Cited Min. Percentile	p value Novelty Def.: 1%	p value Novelty Def.: 10%
Immunology	1985	1%	† 0.000	† 0.000
Immunology	1985	2%	† 0.000	† 0.000
Immunology	1985	5%	† 0.000	**0.000**
Immunology	1985	10%	**0.000**	**0.000**
Immunology	1995	1%	† 0.000	† 0.000
Immunology	1995	2%	† 0.000	† 0.000
Immunology	1995	5%	† 0.000	**0.000**
Immunology	1995	10%	**0.000**	**0.000**
Immunology	2005	1%	† 0.000	† 0.000
Immunology	2005	2%	† 0.000	† 0.000
Immunology	2005	5%	† 0.000	**0.000**
Immunology	2005	10%	**0.000**	**0.000**
Metabolism	1985	1%	**0.000**	**0.000**
Metabolism	1985	2%	**0.000**	**0.000**
Metabolism	1985	5%	**0.000**	**0.000**
Metabolism	1985	10%	**0.000**	**0.000**
Metabolism	1995	1%	**0.000**	**0.000**
Metabolism	1995	2%	**0.000**	**0.000**
Metabolism	1995	5%	**0.000**	**0.000**
Metabolism	1995	10%	**0.000**	**0.000**
Metabolism	2005	1%	† 0.000	**0.000**
Metabolism	2005	2%	**0.000**	**0.000**
Metabolism	2005	5%	**0.000**	**0.000**
Metabolism	2005	10%	**0.000**	**0.000**
Applied Physics	1985	1%	† 0.027	† 0.025
Applied Physics	1985	2%	† 0.000	† 0.000
Applied Physics	1985	5%	**0.000**	**0.000**
Applied Physics	1985	10%	**0.000**	**0.000**
Applied Physics	1995	1%	† 0.010	† 0.013
Applied Physics	1995	2%	**0.000**	**0.000**
Applied Physics	1995	5%	**0.000**	**0.000**
Applied Physics	1995	10%	**0.000**	**0.000**
Applied Physics	2005	1%	† 0.000	**0.000**
Applied Physics	2005	2%	**0.000**	**0.000**
Applied Physics	2005	5%	**0.000**	**0.000**
Applied Physics	2005	10%	**0.000**	**0.000**
Web of Science	1985	1%	**0.000**	**0.000**
Web of Science	1985	2%	**0.000**	**0.000**
Web of Science	1985	5%	**0.000**	**0.000**
Web of Science	1985	10%	**0.000**	**0.000**
Web of Science	1995	1%	**0.000**	**0.000**
Web of Science	1995	2%	**0.000**	**0.000**
Web of Science	1995	5%	**0.000**	**0.000**
Web of Science	1995	10%	**0.000**	**0.000**
Web of Science	2005	1%	**0.000**	**0.000**
Web of Science	2005	2%	**0.000**	**0.000**
Web of Science	2005	5%	**0.000**	**0.000**
Web of Science	2005	10%	**0.000**	**0.000**

These are hypothesis test data for the null hypothesis that hits are distributed among the categories randomly in proportion to the number of articles in each category using a Chi Square Goodness of Fit Test for novel articles defined as those with the 10th percentile z-score being negative and the 1st percentile z-score being negative. Rejecting the null hypothesis supports the alternate hypothesis that hit rates vary among the categories. The p values indicate the significance of the difference between the observed number of hits and the expected number of hits given by the random null model. † denotes that the Chi Square Goodness of Fit Test is not valid because the expected number of hits in at least one category was less than the required five hits. This was caused by the combination of a small number of articles in a category due to a low overall hit rate (1% or 2%) and a definition of novelty (1%) that resulted in few articles being defined as being of low novelty. Results that are significant at the 0.05 level are shown in bold font and those significant at the 0.10 level are shown in italics.

Table S7. Hit Rates By Category

Data Set	Year	Highly Cited Min. Percentile	LNLC	LNHC	HNLC	HNHC
Immunology	1985	1%	0.0	**1.5**	0.6	**1.4**
Immunology	1985	2%	0.0	**3.0**	1.2	**2.8**
Immunology	1985	5%	1.5	6.5	3.2	**7.1**
Immunology	1985	10%	3.0	12.0	7.1	**14.0**
Immunology	1995	1%	0.0	**1.9**	0.5	1.4
Immunology	1995	2%	0.0	**3.4**	1.1	2.8
Immunology	1995	5%	0.6	**7.2**	3.2	6.8
Immunology	1995	10%	1.7	**12.8**	7.6	**12.9**
Immunology	2005	1%	0.0	**1.5**	0.6	**1.3**
Immunology	2005	2%	0.0	**2.7**	1.3	**2.7**
Immunology	2005	5%	0.0	6.0	3.6	**6.7**
Immunology	2005	10%	2.0	10.8	8.1	**12.9**
Metabolism	1985	1%	0.1	**1.5**	0.5	**1.5**
Metabolism	1985	2%	0.3	**2.7**	1.3	**2.9**
Metabolism	1985	5%	0.7	6.6	3.4	**7.0**
Metabolism	1985	10%	2.3	12.3	7.2	**13.8**
Metabolism	1995	1%	0.1	**1.7**	0.6	**1.4**
Metabolism	1995	2%	0.3	**3.1**	1.2	**2.8**
Metabolism	1995	5%	0.7	**7.0**	3.4	**6.7**
Metabolism	1995	10%	1.9	**13.0**	7.4	**13.3**
Metabolism	2005	1%	0.6	**1.3**	0.7	**1.3**
Metabolism	2005	2%	1.0	**2.5**	1.5	**2.6**
Metabolism	2005	5%	2.3	6.0	3.9	**6.6**
Metabolism	2005	10%	4.1	11.4	8.1	**12.6**
Applied Physics	1985	1%	0.0	0.5	**1.2**	**1.2**
Applied Physics	1985	2%	0.9	0.9	**2.5**	**2.4**
Applied Physics	1985	5%	2.8	3.0	5.5	**6.5**
Applied Physics	1985	10%	5.2	6.7	10.6	**13.2**
Applied Physics	1995	1%	0.2	0.7	**1.2**	**1.0**
Applied Physics	1995	2%	0.2	1.3	**2.5**	2.1
Applied Physics	1995	5%	0.9	3.4	**6.0**	5.2
Applied Physics	1995	10%	4.7	7.9	**12.3**	10.9
Applied Physics	2005	1%	0.8	0.6	**1.1**	**1.3**
Applied Physics	2005	2%	1.1	1.3	**2.1**	**2.3**
Applied Physics	2005	5%	1.6	3.3	5.4	**5.9**
Applied Physics	2005	10%	3.9	7.5	10.7	**11.4**
Web of Science	1985	1%	0.4	1.2	1.0	**1.6**
Web of Science	1985	2%	0.9	2.4	2.0	**3.3**
Web of Science	1985	5%	2.6	5.9	5.1	**8.4**
Web of Science	1985	10%	5.8	11.4	10.4	**15.8**
Web of Science	1995	1%	0.4	1.3	0.9	**1.7**
Web of Science	1995	2%	0.9	2.4	1.9	**3.3**
Web of Science	1995	5%	2.5	6.0	5.0	**8.0**
Web of Science	1995	10%	5.6	11.5	10.4	**15.6**
Web of Science	2005	1%	0.4	1.2	1.0	**1.7**
Web of Science	2005	2%	0.9	2.3	2.0	**3.4**
Web of Science	2005	5%	2.5	5.7	5.3	**8.1**
Web of Science	2005	10%	5.6	11.2	10.8	**15.0**

The hit rate is the percentage of publications in the referenced category that are in the top 1%, 2%, 5%, or 10% of papers according to citation count (see column 3) for novel articles defined as those with the 10th percentile z-score being negative. The z-scores are computed using the local network. The category with the highest percentile is boldfaced (the second highest is also boldfaced if within 0.3% and greater than the overall percentage of articles considered to be hits). We also evaluated novelty defined as the 1st percentile of z-scores being negative. We report here on novelty defined at the most stringent parameter setting, while the remaining results are reported at https://bit.ly/2CFMOuf.

Table S8. Explanatory Power of Novelty and Conventionality

Data		Highly Cited	Cumulative Probabilities			
			Conventionality		Novelty	
Set	Year	Min. Percentile	Low	High	Low	High
Immunology	1985	1%	1.000	**0.000**	**0.020**	0.866
Immunology	1985	2%	1.000	**0.000**	**0.002**	0.940
Immunology	1985	5%	1.000	**0.000**	**0.002**	0.924
Immunology	1985	10%	1.000	**0.000**	**0.017**	0.853
Immunology	1995	1%	1.000	**0.000**	**0.000**	0.985
Immunology	1995	2%	1.000	**0.000**	**0.000**	0.991
Immunology	1995	5%	1.000	**0.000**	**0.000**	0.988
Immunology	1995	10%	1.000	**0.000**	**0.000**	0.972
Immunology	2005	1%	1.000	**0.000**	**0.005**	0.882
Immunology	2005	2%	1.000	**0.000**	**0.007**	0.862
Immunology	2005	5%	1.000	**0.000**	**0.012**	0.837
Immunology	2005	10%	1.000	**0.000**	0.265	0.609
Metabolism	1985	1%	1.000	**0.000**	**0.000**	0.991
Metabolism	1985	2%	1.000	**0.000**	**0.000**	0.986
Metabolism	1985	5%	1.000	**0.000**	**0.000**	0.999
Metabolism	1985	10%	1.000	**0.000**	**0.000**	0.997
Metabolism	1995	1%	1.000	**0.000**	**0.000**	1.000
Metabolism	1995	2%	1.000	**0.000**	**0.000**	1.000
Metabolism	1995	5%	1.000	**0.000**	**0.000**	1.000
Metabolism	1995	10%	1.000	**0.000**	**0.000**	1.000
Metabolism	2005	1%	1.000	**0.000**	**0.000**	0.997
Metabolism	2005	2%	1.000	**0.000**	**0.000**	0.998
Metabolism	2005	5%	1.000	**0.000**	**0.000**	0.998
Metabolism	2005	10%	1.000	**0.000**	**0.000**	0.996
Applied Physics	1985	1%	0.177	0.860	0.998	0.071
Applied Physics	1985	2%	0.066	0.952	1.000	**0.013**
Applied Physics	1985	5%	0.200	0.818	1.000	**0.003**
Applied Physics	1985	10%	0.390	0.625	1.000	**0.000**
Applied Physics	1995	1%	0.062	0.955	0.996	0.090
Applied Physics	1995	2%	**0.018**	0.988	1.000	**0.011**
Applied Physics	1995	5%	**0.002**	0.999	1.000	**0.001**
Applied Physics	1995	10%	**0.000**	1.000	1.000	**0.000**
Applied Physics	2005	1%	0.319	0.706	1.000	*0.028*
Applied Physics	2005	2%	0.272	0.748	1.000	**0.015**
Applied Physics	2005	5%	0.117	0.897	1.000	**0.000**
Applied Physics	2005	10%	0.102	0.909	1.000	**0.000**
Web of Science	1985	1%	1.000	**0.000**	0.986	**0.002**
Web of Science	1985	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	1985	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	1985	10%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	1%	1.000	**0.000**	0.969	**0.007**
Web of Science	1995	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	1995	10%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	1%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	2%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	5%	1.000	**0.000**	1.000	**0.000**
Web of Science	2005	10%	1.000	**0.000**	1.000	**0.000**

This table lists p-values in the form of cumulative right-hand tail probabilities for the observed number of hits in the Low Novelty, High Novelty, Low Conventionality, and High Conventionality categories under the sampling distribution generated by the null hypothesis of a random distribution of hit articles in proportion to the number of articles in each of the categories. A small p-value, therefore, indicates a number of hits that exceeds the expected number. Results that indicate statistically significant numbers of hits in excess of the expected number at the 0.05 level using a two-tailed test are highlighted in bold font, and those significant at the 0.10 level are italicized. These data are for the circumstances where novel citation patterns are defined by whether an article's 10th percentile z-score is negative. The z-scores are computed using the local network. We report these data because this is the most stringent definition of novelty. We also report results for novel articles defined by the 1st percentile at https://bit.ly/2CFMOuf.

References

1. Korobskiy D, Davey A, Liu S, Devarakonda S, Chacko G (2019) Enhanced Research Network Informatics Environment (ERNIE), (NET ESolutions Corporation), Github repository. https://github.com/NETESOLUTIONS/ERNIE/tree/master/P2_studies/Permutation_Testing.
2. Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science (New York, N.Y.)* 342(6157):468–472.